Speech and Voice Recognition White Paper

This white paper differentiates between speech recognition and speaker/voice recognition and provides a basic analysis of respective market size.

Rawlson O`Neil King
Lead Researcher, Biometrics Research Group, Inc.

VOICETRUST

Leaders in Voice Biometrics

- Enhanced Security
- Convenient Authentication
- Significant Cost Savings

TABLE OF CONTENTS

Research Methodology

Biometrics Research Group, Inc. uses a combination of primary and secondary research methodologies to compile the necessary information for its research projections.

The conclusions drawn are based on our best judgment of exhibited trends, the expected direction the industry may follow, and consideration of a host of industry drivers, restraints, and challenges that represent the possibility for such trends to occur over a specific time frame. All supporting analyses and data are provided to the best of ability.

Primary Research

Biometrics Research Group, Inc. conducts interviews with technology providers, clients, and other organizations, as well as stakeholders in each of the technology segments, standards organizations, privacy commissions, and other influential agencies. To provide balance to these interviews, industry thought leaders who track the implementation of the biometric technologies are also interviewed to get their perspective on the issues of market acceptance and future direction of the industry.

Biometrics Research Group, Inc. also applies its own proprietary micro- and macroeconomic modeling using a regression analysis methodology to determine the size of biometric and related-industry marketplaces. Using databases of both publicly and privately-available financial data, Biometrics Research Group works to project market size and market potential, in the context of the global economic marketplace, using proven econometric models.

Secondary Research

Biometrics Research Group, Inc. also draws upon secondary research which includes published sources such as those from government bodies, think tanks, industry associations, internet sources, and Biometrics Research Group, Inc.'s own repository of news items. This information was used to enrich and externalize the primary data. Data sources are cited where applicable.

Speaker/Voice Recognition and Speech Recognition Differentiation

Speaker, or voice, recognition is a biometric modality that uses an individual's voice for recognition purposes. It is described as a process by which a machine or program receives and interprets dictation as well as understands and carries out spoken commands. Voice recognition technologies help customers comply with privacy, security, and safety requirements governed by law and by user expectations. It is a different technology than "speech recognition," which recognizes words as they are articulated, which is not a biometric.

Speaker or voice recognition processes rely on features influenced by both the physical structure of an individual's vocal tract and the behavioral characteristics of the individual. Speaker recognition has been applied most often as a security application to control access to buildings or sensitive data. Banking and financial institutions have employed speaker verification as a security mechanism on telephone-initiated transfers of large sums of money. In addition to adding security, verification is advantageous because it reduces the turnaround time on banking transactions. Speaker verification can also be used by firms to limit data access to authorized personnel. Speaker recognition also provides a mechanism to limit the remote access of a personal workstation to its owner or a set of registered users. In addition to its use as a security device, speaker recognition can be used to trigger specialized services based on a user's identity.

Speech Recognition

Speech recognition, in contrast, is most often applied in manufacturing for companies needing voice entry of data or commands while the operator's hands are otherwise occupied. Related applications occur in product inspection, inventory control, command/control, and material handling. Speech recognition also finds frequent application in medicine, where voice input can significantly accelerate the writing of routine reports. Furthermore, speech recognition helps users control personal workstations or interact with other applications remotely when touch-tone keypads are not available.

Most systems for speech recognition include the following five components:

1. A speech capture device or input. This usually consists of a microphone and associated analog-to-digital converter, which digitally encodes the raw speech waveform.

2. A digital signal processing module. The DSP module performs endpoint (word boundary) detection to separate speech from non-speech, converts the raw waveform into a frequency domain representation, and performs further windowing, scaling, filtering, and data compression. The goal is to enhance and retain only those components of the spectral representation that are useful for recognition purposes, thereby reducing the amount of information that the pattern-matching algorithm must contend with. A set of these speech parameters for one interval of time (usually 10-30 milliseconds) is called a speech frame.

3. Preprocessed signal storage. Here, the preprocessed speech is buffered for the recognition algorithm.

4. Reference speech patterns. Stored reference patterns can be matched against the user's speech sample once it has been preprocessed by the DSP module. This information is stored as a set of speech templates or as generative speech models.

5. A pattern matching algorithm. The algorithm must compute a measure of goodness-of-fit between the preprocessed signal from the user's speech and all the stored templates or speech models. A selection process chooses the template or model (possibly more than one) with the best match.

Five factors can be used to control and simplify the speech recognition task:

1. Isolated words. Speech consisting of isolated words (short silences between
the words) is much easier to recognize than continuous speech because word boundaries are difficult to

find in continuous speech. Also, "co-articulation" effects in continuous speech cause the pronunciation of a word to change depending on its position relative to other words in a sentence. Error rates can be reduced by requiring the user to pause between each word. However, this type of restriction places a burden on the user and reduces the speed with which information can be input to the system.

2. Single speaker. Speech from a single speaker is also easier to recognize than speech from a variety of speakers because most parametric representations of speech are sensitive to the characteristics of the particular speaker. This makes a set of pattern-matching templates for one speaker perform poorly for another speaker. Therefore, many systems are speaker dependent - trained for use with each different operator. Relatively few speech recognition systems can be used by the general public. A rule of thumb used by many researchers is that for the same task, speaker-dependent systems will have error rates roughly three to five times smaller than speaker-independent ones. One way to make a system speaker independent is simply to mix training templates from a wide variety of speakers. A more sophisticated approach will attempt to look for phonetic features that are relatively invariant between speakers.

3. Vocabulary size. The size of the vocabulary of words to be recognized also strongly influences recognition accuracy. Large vocabularies are more likely to contain ambiguous words than small vocabularies. Ambiguous words are those whose pattern-matching templates appear similar to the classification algorithm used by the recognizer. They are therefore harder to distinguish from each other. Of course, small vocabularies composed of many ambiguous words can be particularly difficult to recognize. The amount of time it takes to search the speech model database also relates to vocabulary size. Systems containing many pattern templates typically require pruning techniques to cut down the computational load of the pattern-matching algorithm. By ignoring potentially useful search paths, pruning heuristics can also introduce recognition errors.

4. Grammar. The grammar of the recognition domain defines the allowable sequences of words. A tightly constrained grammar is one in which the number of words that can legally follow any given word is small. The amount of constraint on word choice is referred to as the perplexity of the grammar. Systems with low perplexity are potentially more accurate than those that give the user more freedom because the system can limit the effective vocabulary and search space to those words that can occur in the current input context.

5. Environment. Background noise, changes in microphone characteristics, and loudness can all dramatically affect recognition accuracy. Many recognition systems are capable of very low error rates as long as the environmental conditions remain quiet and controlled. However, performance degrades when noise is introduced or when conditions differ from the training session used to build the reference templates. To compensate, the user most usually has to always wear a headset-mounted, noise-limiting microphone.

Speaker or Voice Recognition

Voice recognition has a history dating back some four decades and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (such as size and shape of the throat and mouth) and learned behavioral patterns (such as voice pitch and speaking style).

There are two major applications between voice recognition technologies and methodologies. If a speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication.

On the other hand, identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas speaker identification is a 1:N match where the voice is compared against N templates.

Voice recognition is a popular choice for remote authentication due to the availability of devices for collecting speech samples, such as telephone systems and computer microphones. Due to its ease of integration, speaker recognition is different from some other biometric methods in that speech samples are captured dynamically or over a period of time, such as a few seconds. Analysis occurs on a model in which changes over time are monitored, which is similar to other behavioral biometrics such as dynamic signature, gait, and keystroke recognition.

The physiological component of voice recognition is related to the physical shape of an individual's vocal tract, which consists of an airway and the soft tissue cavities from which vocal sounds originate. To produce speech, these components work in combination with the physical movement of the jaw, tongue, and larynx and resonances in the nasal passages. The acoustic patterns of speech come from the physical characteristics of the airways.

Motion of the mouth and pronunciations are the behavioral components of this biometric. There are two forms of speaker recognition: text dependent known as "constrained mode" and text independent known as "unconstrained mode".

In a system using "text dependent" speech, the individual presents either a fixed or prompted phrase that is programmed into the system and can improve performance especially with cooperative users.

A "text independent" system has no advance knowledge of the presenter's phrasing and is much more flexible in situations where the individual submitting the sample may be unaware of the collection or unwilling to cooperate, which presents a more difficult challenge. Speech samples are waveforms with time on the horizontal axis and loudness on the vertical access. The speaker recognition system analyzes the frequency content of the speech and compares characteristics such as the quality, duration, intensity dynamics, and pitch of the signal.

In "text dependent" systems, during the collection or enrollment phase, the individual says a short word or phrase -- referred to as an utterance -- typically captured using a microphone that can be as simple as a telephone. The voice sample is converted from an analog format to a digital format, the features of the individual's voice are extracted, and then a model is created.

Most "text dependent" speaker verification systems use the concept of Hidden Markov Models (HMMs), random based models that provide a statistical representation of the sounds produced by the individual. The HMM represents the underlying variations and temporal changes over time found in the speech states using the quality duration / intensity dynamics / pitch characteristics mentioned above.

Another method is the Gaussian Mixture Model, a state-mapping model closely related to HMM, that is often used for unconstrained "text independent" applications. Like HMM, this method uses the voice to create a number of vector "states" representing the various sound forms, which are characteristic of the physiology and behavior of the individual.

These methods all compare the similarities and differences between the input voice and the stored voice "states" to produce a recognition decision. After enrollment, during the recognition phase, the same quality / duration / loudness / pitch features are extracted from the submitted sample and compared to the model of the claimed or hypothesized identity and to models from other speakers.

The other-speaker or "anti-speaker" models contain the "states" of a variety of individuals, not including that of the claimed or hypothesized identity. The input voice sample and enrolled models are

compared to produce a "likelihood ratio," indicating the likelihood that the input sample came from the claimed or hypothesized speaker. If the voice input belongs to the identity claimed or hypothesized, the score will reflect the sample to be more similar to the claimed or hypothesized identity's model than to the "anti-speaker" model.

The seemingly easy implementation of speaker recognition systems contributes to the process major weakness and susceptibility to transmission channel and microphone variability and noise.

Systems can face problems when end users have enrolled on a clean landline phone and attempt verification using a noisy cellular phone. The inability to control the factors affecting the input system can significantly decrease performance. Speaker verification systems, except those using prompted phrases, are also susceptible to spoofing attacks through the use of recorded voice. Anti-spoofing measures that require the utterance of a specified and random word or phrase are being implemented to combat this weakness.

For example, a system may request a randomly generated phrase, to prevent an attack from a pre-recorded voice sample. The user cannot anticipate the random sample that will be required and therefore cannot successfully attempt a "playback" spoofing attack on the system.

Current research in the area of "text independent" speaker recognition is mainly focused on moving beyond low-level spectral analysis. Although the spectral level of information is still the driving force behind the recognitions, fusing higher-level characteristics with the low level spectral information is becoming a popular laboratory technique.

Speaker recognition characteristics such as rhythm, speed, modulation and intonation are based on personality type and parental influence; and semantics, idiolects, pronunciations and idiosyncrasies are related to birthplace, socio-economic status, and education level. Higher-level characteristics can be combined with the underlying low-level spectral information to improve the performance of "text independent" speaker recognition systems.

The great advantage of speaker verification is its widespread acceptability and ease of use, as well as the relative inexpensiveness of basic systems compared to other biometric options. When voice authentication is integrated into a telephony system, it also creates a very friendly customer service environment. As a consequence, respondents in numerous surveys have indicated that they prefer recognition technology for biometric identification. Voice recognition is characterized by non-contact and non-intrusiveness. Due to ease of use, voice recognition is a growing market segment, especially in the financial sector.

Financial institutions have identified voice biometrics as one of the best means to secure its client accounts and financial information. Voice biometrics compares various characteristics drawn from a person's voice such as inflection, pitch, dialect, among others, and matches that with data captured. For voice recognition to work it requires banks and other financial institutions to register their clients voice patterns and correlate them to personal data for incorporation into a database.

Voice biometrics solutions allow customers to verify their identity simply by speaking, making it easier and faster to gain access to secure banking and insurance services by way of mobile apps, telephone and Web channels. Voice biometric solutions eliminate the need for PIN-based password or interrogation-based authentication methods, or can be used to add another level of security to those systems.

Banks that deploy voice biometrics to automate the "login" process not only enhance customer satisfaction levels, but dramatically reduce their customer care costs through increased automation rates.

Due to the versatility, along with consumer confidence in voice biometric technology, the Biometrics Research Group expects voice biometrics to be the fasting growing technology modality in the banking sector.

Surveys we have analyzed have found that consumers prefer voice recognition technology for biometric identification. According to a survey conducted by IT provider Unisys, the biometric modalities ranked by consumer preference are: voice recognition (32 percent), fingerprints (27 percent), facial scan (20 percent), hand geometry (12 percent), and iris scan (10 percent). As a result, the Biometrics Research Group projects that voice recognition will be widely adopted. We project the technology will not only be implemented in bank calling centers throughout the world, but fast growth will also be driven by the continued rapid worldwide adoption of mobile "smartphone" and "superphone" technologies.

Banks are in the preliminary stages of testing and rolling out new voice biometric technologies for mobile devices. In North America last year, USAAM, the independent bank and insurance brokerage that caters to the U.S. military, developed a voice recognition service that will eventually allow its entire mobile phone customer base to make natural language inquiries for a wide range of banking services.

USAA's voice recognition app is currently being tested by a group of employees but is slated for use by the bank's 6.3 million account holders early next year. The bank has stated publicly that the application has tremendous potential to make banking: "simpler, faster and more satisfying on mobile devices."

The bank cites the fact that military personnel are often quite mobile and would like to make greater use of advanced smartphone and superphone technologies. The bank also cited statistics from its voice biometric technology supplier Nuance that while over 50 percent of smartphone and superphone owners have installed a mobile banking app on their device, only 27 percent actually use it on a regular basis. Consumers say improvements in a few areas would increase the use of smartphone and superphone banking apps significantly. Thirty-four percent say they would appreciate a seamless access to a live agent when they need one, and 21 percent want their mobile apps to include more self-service tasks. Eighteen percent would simply settle for an app that was easier to use.

Technology suppliers like Nuance are betting that voice biometrics will be the "magic sauce" that improves banking customer experiences. Other banks are also examining implementation of the voice biometrics technology. Spanish bank BBVA is also currently developing a Siri-like banking application for iPhones and iPads at its U.S. subsidiary.

Further, as BiometricUpdate.com previously reported last year, major banks such as ANZ are seriously beginning to study implementing biometrics over the next three to five years to improve the quality of its banking services. ANZ has made public statements that it projects it will take two to three years before commercialization of biometrics in banking is achieved. However, the bank is positioning itself for the implementation of the new technologies that will be designed to simplify ANZ's distribution networks and its products and processes, while providing customers with additional mobile and flexible banking options, while concurrently improving the capability of front-line staff.

Market Size

The market size for both speech recognition and voice recognition continue to grow concurrently. While many research firms do not differentiate between the speech recognition and speaker recognition marketplaces in their analysis, the Biometrics Research Group does.

We recognize that various industries such as airlines, banks, and brokerages depend on voice recognition functionality, not only to enhance their customer contacts, but also to comply with security requirements dictated by the law and the security conscious expectations of customers. As a result, voice recognition as a biometric modality has been growing. Biometrics Research Group, Inc. estimates that voice recognition will reach US$2.5 billion in revenue by 2015, mainly driven by the banking sector.

Due to growing interest in providing consumers with cutting edge technology, while concurrently enhancing banking security, Biometrics Research Group expects more financial institutions to develop and deploy biometrics, and as a consequence, expects revenue growth for voice biometrics to grow. Our research estimates that at least US$200 million was spent on voice biometrics in the banking sector in 2012. We estimate that at least US$750 million will be spent on voice biometrics in the banking sector by 2015.

While we estimate that growth will continue to occur, its pace will be dictated by technological developments. One of the major factors that we have determined is restraining the voice recognition marketplace are system faults that cause false inputs produced by poor communication linkages, loud external conversations, barking dogs, screaming children and the like. As these system errors begin to be addressed by technological innovation, the Biometrics Research Group anticipates exponential increases in year-over-year earnings for vendors in the voice recognition space.

Biometrics Research Group also expects that the increased use of mobile devices will also drive voice recognition development. Currently, security measures to lock smartphones, and the data contained within, include: four-digit passcodes, draw pattern unlock algorithms and increasingly fingerprint impressions. The most obvious and conceivable security measure for mobile devices however is voice recognition. We would expect that once voice recognition system error issues are resolved, major smartphone manufacturers will move towards wholesale implementation of the technology for device lock purposes. Such a move will expand compound annual growth rates and increase total market value, though we cannot estimate when these developments might occur.

In differentiating the marketplace, Biometrics Research Group notes that the speech recognition market is much larger and more mature. We estimate speech recognition software sales were valued at US$11.5 billion in 2010 and will reach US$20.1 billion in 2015. Software packages include automatic speech recognition and text-to-speech systems. It should be noted that because we do not consider speech recognition itself a biometric, we do not include revenue projections from this segment in our global biometric revenue estimates. Due to vendor crossover in concurrent markets however we do singularly identify main champions in the "unified space". Biometrics Research Group believes that the main vendors controlling the space, based on market share, include: Nuance Communications, ValidSoft Ltd, and VoiceTrust.

Nuance Communications

Nuance Communications is an American multinational computer software technology corporation, headquartered in Burlington, Massachusetts that provides speech and imaging applications. Current business products focus on server and embedded speech recognition, telephone call steering systems, automated telephone directory services, medical transcription software & systems, optical character recognition software, and desktop imaging software. The company also maintains a small division which does software and system development for military and government agencies.

Nuance was founded in 1994 as a spin-off of SRI International's Speech Technology and Research

(STAR) Laboratory to commercialize the speaker-independent speech recognition technology developed for the U.S. government at SRI. Initially based in Menlo Park, California, Nuance deployed their first commercial large-scale speech application in 1996. Their initial route to market was through call center automation. Call centers had just centralized the branch-office telephone handling function throughout many large companies. The highest cost of running call centers is the cost of staff. Early projects were completely developed by Nuance to prove commercial practicality and provide efficiency benefits.

After formation, the company expanded as a result of organic growth, mergers, and acquisitions. In October 2005, Nuance merged with ScanSoft, a fellow commercial large scale speech application business. ScanSoft – a Xerox spin-off – had its roots in Kurzweil Computer Products, a software company that developed the first omni-font character recognition system. In October 2011, unconfirmed research suggested that Nuance's servers power Apple's iPhone Siri voice recognition application, reaffirming the firm's tradition of innovation. The firm is best known in the consumer marketplace for Dragon, the world's bestselling speech recognition software for Macintosh and PC.

ValidSoft

ValidSoft is a UK-based security software company, providing telecommunications-based multi-factor authentication, identity and transaction verification technology. The company was founded in 2003 and provides business-to-business mobile security and cloud security products, including a multi-factor authentication platform called SMART (Secure Mobile Architecture for Real-time Transactions), which uses mobile telecommunication channels and devices and includes a proprietary voice biometric engine. These solutions are designed to reduce electronic fraud and safeguard consumer privacy when using Internet and mobile banking, credit, debit card and both mobile and fixed line telephony channels. ValidSoft's products are designed to verify the authenticity of both parties to a transaction (mutual authentication), ensure the fidelity of telecommunication channels (secure communications), and confirm the integrity of transactions themselves (transaction verification).

VALid-SVP, otherwise known as the VALid Speaker Verification Platform, is ValidSoft's proprietary voice biometrics solution, based on a modular and pluggable architecture, which allows organisations to easily integrate voice verification into a broader authentication platform. VALid-SVP is described as a leading-edge voice biometrics engine that supports text-dependent, text-independent and conversational voice verification (biometric plus knowledge) models.

It includes a full enrolment module, incorporates "liveness" validation, replay and synthesis attack mitigation techniques and "pseudo device theft" detection as well as many other leading security techniques as part of its extensive layered security model. It operates on any channel the VALid platform supports, including Internet, mobile, IVR and contact-center and can also be used cross-channel. Additionally, VALid-SVP also utilizes more advanced techniques to ensure the integrity of voice verification in real-world usage. This includes techniques such as context aware voice verification and dynamic threshold adjustments. As the VALid platform understand many facets of the overall authentication all of these aspects can be considered when a conclusion is reached. The firm's aim is to allow the user to transact securely in the manner they wish to operate within in the most frictionless way.

ValidSoft's VALid-IMA is described as an "in-band" mobile authentication solution because the primary channel of voice communication between the app and the IMA Server is not actually the voice channel, i.e. through a phone call to/from the handset, but via the data channel on which

the app it is integrated with is communicating to its own host. In-band voice delivery, coupled with the functionality of the IMA platform and the advanced features of ValidSoft's Voice Biometric engine, delivers a combination of service delivery cost reduction, low user friction, biometric performance improvement, secure and simple enrolment/activation of the app and user-friendly features that can eliminate false-negatives, i.e. denial of access to a legitimate customer. VALid-IMA can be integrated into apps on any operating system and on any handset because it simply uses a well-defined XML protocol for communication with the IMA Server. It can also be deployed as either an in-house implementation or software-as-a-service.

VoiceTrust

VoiceTrust is a global provider of voice biometrics solutions that enable highly secure and convenient authentication. The world's largest banks, insurance companies, call centers, and enterprises rely on the firm's solutions to protect access to business and consumer applications, prevent identity theft, and deliver a more enjoyable authentication experience. Founded in 2000, the privately-owned company is headquartered in Toronto, Canada with additional offices in the U.S., Germany, Netherlands, and UAE.

The company's voice biometrics platform, known as VTAssure, can verify identity based on text dependent and text independent authentication. This authentication platform consists of both an interactive call management platform and voice verification platform, using a comprehensive multifactor authentication algorithm and stored biometric templates. VoiceTrust also recently released VT inTalk, a new solution for dynamic caller verification. VT inTalk is an off-the-shelf, fast deployable solution that addresses the needs of call centers wanting to verify the authenticity of a caller while engaged in natural dialogue with the customer. VT inTalk dynamically compares the speaker's voice with the existing claimant voice print to provide additional assurance that the caller is who he or she claims to be. The solution will be integrated into the Genesys GVP platform. The Genesys Interactive Voice Response Platform is a software only, standards-based voice portal that enables businesses to provide cost-effective customer interactions 24×7 for voice, video, and web-based interactions. Beyond traditional interactive voice response (IVR) systems, it provides touchtone access to applications and incorporates speech recognition technology and video for conversational exchange to identify and resolve customer requests.

VoiceTrust's other recent clients and pilot projects include a major U.S.-based bank, an Irish bank, a global U.S.-based apparel manufacturer, a European insurance provider, a European government agency as well as several financial service providers and a telecommunication firm in the Middle East.

All of the above vendors are attempting to dominate the combined speech and voice recognition marketplaces by providing differentiated products and services designed to provide a competitive edge to their clients. Biometrics Research Group expects more stiff competition between these vendors as demand for voice recognition systems grows within the banking and mobile sectors.

About the Biometrics Research Group, Inc.

Biometrics Research Group, Inc. provides proprietary research, consumer and business data, custom consulting, and industry intelligence to help companies make informed business decisions.

We provide news, research and analysis to companies ranging from Fortune 500 to small start-ups through market reports, primary studies, consumer research, custom research, consultation, workshops, executive conferences and our free daily BiometricUpdate.com news service.

Biometrics Research Group has positioned itself as the world's preferred supplier of pure-play market research and consultancy services focused on the biometric marketplace, which particular focus on the law enforcement and national security sectors. Our portfolio of white papers and full research reports is based upon high-quality quantitative analysis, allowing our clients to gain deeper understanding of the marketplace.

We customize our research design, data collection, and statistical reporting using proprietary micro- and macroeconomic modeling and regression analysis.

Through integration of our research results with qualitative analysis from our BiometricUpdate.com news service, we provide actionable business analysis.